Revisiting Classifier: Transferring Vision-Language Models for Video Recognition

Wenhao Wu1,2, Zhun Sun2, Wanli Ouyang1,3

1The University of Sydney 2Baidu Inc. 3Shanghai AI Laboratory

MOTIVATION

Observation: the semantic information contained in the samples may correlate with inter-classes.

CONTRIBUTION

We build a new recognition paradigm to improve the transferability using visual knowledge and textual knowledge from the well-pre-trained vision-language model.

We conduct extensive experiments on popular video datasets (i.e., Kinetics-400 & 600, UCF-101, HMDB51 and ActivityNet) to demonstrate the transferability of our solution in many types of transfer learning, i.e., zero-shot / few-shot / general video recognition. Our approach democratizes the training on video datasets and achieves state-of-the-art performance on various video recognition settings, e.g., 87.8% top-1 accuracy on Kinetics-400, and outperforms previous methods by 20∼50% absolute top-1 accuracy under zero-shot, few-shot settings.

METHOD

Revisiting Classifier: From a frozen classifier perspective

Q: How to obtain inter-class correlation?

A1: Transferring visual statistical knowledge.

A2: Transferring textual semantic knowledge.

ABLATION STUDIES

Comparison with vision-only framework

Comparison with contrastive-based framework

Results on Kinetics-400 dataset

Comparisons withFew-shot SOTAs

Comparison with Zero-shot SOTAs

Method

UCF101 / HMDB51

Method

Top-1 mAP

Results on Kinetics-400 dataset

Method

Top-1 mAP

Comparison with SOTAs

Method

Input Pretrained Top-1 Top-5 FLOPs/View Params

Results on ActivityNet

Method

UCF101 HMDB51

Method

Top-1 mAP

Comparison with Few-shot SOTAs

Method

shot HMDB UCF ANet K400

Comparison with SOTAs

Method

UCF1 / UCF50 HMDB51 / HMDB51 Anet/UCF Anet

Method

Comparison with Zero-shot SOTAs

Method

UCF / UCF50 HMDB51 / HMDB51 Anet/UCF Anet

Method

Comparison with UCF / UCF50 HMDB51 / HMDB51 Anet/UCF Anet

Comparison with Zero-shot SOTAs

Method

Zero-Train / Few-Train / Zero-Test

Method

Comparison with Few-shot SOTAS

Method

VideoSwin [36] 2 20.9 33.3 -

VideoPrompt [25] 5 56.6 79.5 -

X-Flavours [40] 2 51.6 84.0 -

Ours ViT/L 1 72.7 96.4 89.0 75.8

Ours ViT/L 1 72.7 96.4 89.0 75.8

Ours ViT/L 1 72.7 96.4 89.0 75.8

Ours ViT/L 1 72.7 96.4 89.0 75.8

Ours ViT/L 1 72.7 96.4 89.0 75.8

Comparison with UCF / UCF50 HMDB51 / HMDB51 Anet/UCF Anet

Method

Comparison with Few-shot SOTAS

Method

VideoSwin [36] 2 20.9 33.3 -

VideoPrompt [25] 5 56.6 79.5 -

X-Flavours [40] 2 51.6 84.0 -

Ours ViT/L 1 72.7 96.4 89.0 75.8

Ours ViT/L 1 72.7 96.4 89.0 75.8

Ours ViT/L 1 72.7 96.4 89.0 75.8

Ours ViT/L 1 72.7 96.4 89.0 75.8

Ours ViT/L 1 72.7 96.4 89.0 75.8

Comparison with SOTAs

Method

Input Pretrained Top-1 Top-5 FLOPs/View Params

Results on ActivityNet

Method

UCF101 HMDB51

Method

Top-1 mAP

Comparison with Few-shot SOTAs

Method

shot HMDB UCF ANet K400

Comparison with SOTAs

Method

UCF101 / HMDB51

Method

Top-1 mAP